Text Entry Specs

(by Kaja Mueller-Wang and Duncan Paterson, Cluster Asia and Europe in a Global Context, Heidelberg University, Germany)

Version 1.1 April 22, 2015

Table of Contents

File Conventions

Save each book as xml file (.xml) with UTF-8 encoding. Please ensure that all xml files are well-formed. The typed texts should appear inside the root element <text></text>.

Example:

```
<?xml version="1.0" encoding= "UTF-8"?>
<text>
   [pdf contents here]
</text>
```

General Markup

Type the contents of each page separately before moving on to the next page.

Page Breaks, Page Numbers and Running Heads

Page breaks are marked by “<pb/>”. Running heads appear either in the left or right margin of the page. Type the running heads immediately before or after page breaks. If there is a printed page number on the page include it as part of the running head. Running heads are marked by <fw type="footer"> for running heads appearing in the left margin of the page and <fw type="header"> for running heads appearing in the right margin of the page. If a page number appears within the running head include it between the <fw> tags. Please type large spaces in the running head as a single IDEOGRAPHIC SPACE character U+3000.

Please note: In the digitization of the book, the two half-pages may be on the same scan or on two consecutive scans.

Example:

```
<fw type="footer">文章辨體序說　一三</fw>
```

Text Blocks

Each section of text that forms a semantic unit on a single or stretching multiple pages should be marked by “ ” element tags. Headers, titles, subtitles, etc. are part of the and should appear inside them.

Headings

Enclose each heading (unless it appears in a running head, s.a.) with <head></head> tags. Please note in the books to be typed headings always appear between elements (s.b.).

Paragraphs

The beginning of a new paragraph is either signaled by a header or by indentation. Paragraphs should be marked by elements.

Line Breaks

Type ”<lb/>” at the end of each line to determine line-breaks.

Example:

A section beginning with a header followed by paragraphs, including line breaks.

```
<div>
  <head>樂府</head><lb/>
  <p>易曰[ …]得錄云。</p><lb/>
</div>
```

Front and Back matter

If prefaces, title pages, table of contents (s.b.), etc, appear in the text they should be enclosed by “<front>” or “<back>” respectively. Front and back matter can appear relative to the whole text, or relative to one of the titles within the volume. If more than one book has been combined within the scans that we submitted, each book forms a “<group>”.

Example: 涵芬樓文談

涵芬樓文談 contains two books in one printed volume with prefatory matter for the whole book, as well as prefatory matter of the two books included in this volume. The book corresponds to the <text> elements. <front> covers the preface to the the whole book. The two titles included in the book are marked as a <group>. Immediately followed by <text>. The prefatory matter of the first book appears inside another <front> element. The main body of the text appears inside a <body> element.

In the more common case where only one title appears inside a book, there is a single pair of <body></body> elements to mark the main text, as opposed to dedications, colophons, etc.

```
<text>
    <front>
      <div>
        <head>校點前言</head><lb/>
        <p>[Some text here.]</p>
      </div>
    </front>
    <group>
      <text>
        <front>
          <head>文章辨體序說</head><lb/>
        </front>
        <body>
          <div>
            <head>[Some text here.]</head>
            <p>[Some text here.]</p>
          </div>
        </body>
      </text>
      <text>
```

```
    <front>
      <head>文體明辨序說</head><lb/>
    </front>
    <body>
      <div><p>[Some text here.]</p></div>
    </body>
  </text>
</group>
</text>
```

Structured Text

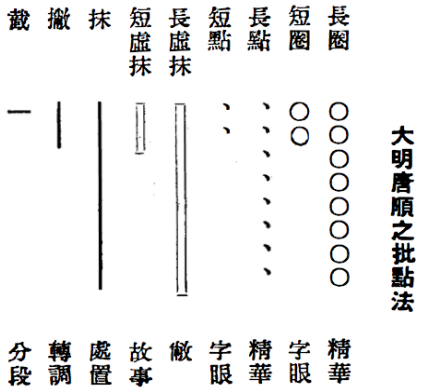
Tables

Tables are enclosed by <table></table> including the number of rows and columns. Include the number of columns as *col="[some number]"* and the number of rows as *row="[some number]"*. The head or title of the table is included as <head> (s.a.) inside the table. The sequence of rows and columns is in the order of reading of the main body of text. (From right top to left bottom in the example below). Line-breaks are omitted inside tables.

Example[1]:
A table containing special symbols.

[1] For list of the symbols in row two see below Special Symbols and Characters.

```
<table rows="3" cols="9">
  <head>大明唐順之批點法</head>
  <row>
    <cell>長圈</cell>
    <cell>短圈</cell>
    <cell>長點</cell>
        .
        .
        .
  </row>
  <row>
    <cell>○○○○○○○○</cell>
    <cell>○○</cell>
    <cell>、、、、、、、、、</cell>
        .
        .
        .
  </row>
  <row>
    <cell>精華</cell>
    <cell>字眼</cell>
    <cell>精華</cell>
        .
        .
        .
  </row>
 </table>
```



Lists

Lists should be embedded within “<list>” tags. Please follow the logical order of the list by nesting the individual items to create new lists. Since we follow the logical order within lists, line-breaks are omitted.

Example:
A generic nested list.

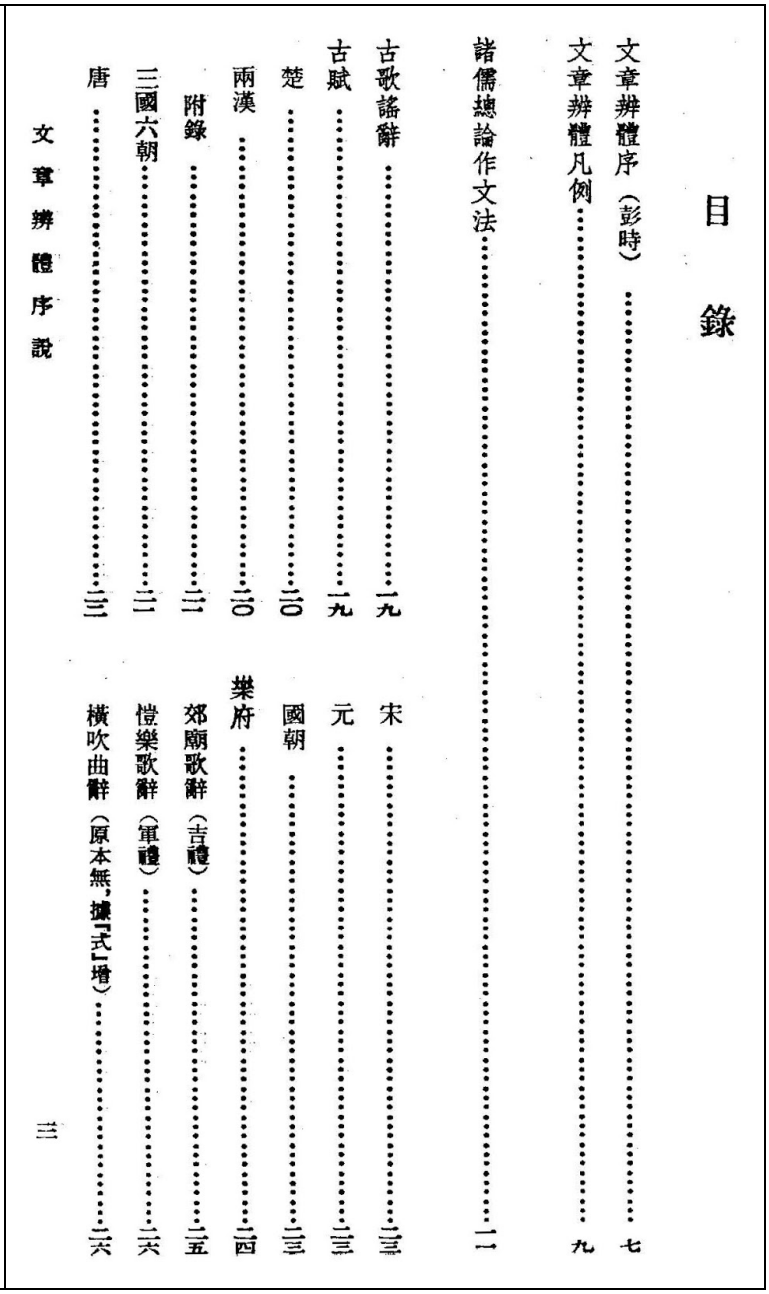
```
<list>
  <item>1</item>
  <item>2
    <list>
      <item>2.1</item>
      <item>2.2</item>
    </list>
  </item>
</list>
```

Table of Contents (目錄)

Tables of Contents are lists, we are not interested in their layout just their logical structure. Sub-sections should appear within their appropriate sections. Use one HORIZONTAL ELLIPSIS "…" U+2026 for dots that fill spaces in lists

Example:
Sample page of Table of Contents.

```
<div type="contents">
  <head>目錄</head>
  <list>
    <item>文章辨體序<hi>（彭時）</hi>…七</item>
    <item>文章辨凡例…九</item>
  </list>
  <list>
    <item>諸儒總論作文法…一一</item>
  </list>
  <list>
    <item>古歌謠詞…一九</item>
    <item>古賦…一九</item>
    <list>
      <item>楚…二〇</item>
      <item>兩漢…二〇</item>
      <list>
        <item>附錄…二一</item>
      </list>
      <item>三国六朝…二一</item>
      <item>唐…二二</item>
      <item>宋…二三</item>
[…]
<fw type="footer">文章辨體序說　三</fw>
<pb/>
[…]
    </list>
  </list>
</div>
```



目錄

文章辨體序（彭時）……七
文章辨體凡例……九
諸儒總論作文法……一一
古歌謠辭……一九
古賦……一九
楚……二〇
兩漢……二〇
附錄……二一
三國六朝……二一
唐……二二
宋……二三
元……二三
國朝……二三
樂府……二四
郊廟歌辭（吉禮）……二五
愷樂歌辭（軍禮）……二六
橫吹曲辭（原本無，據「式」增）……二六

文章辨體序說　三

Special Symbols and Characters

Character variants and unknown characters

If a character variant is included in Unicode 7.0, type it. Do not normalize the variant.

Example:
If the character variant 歷 of the character 歷 occurs in the text, type 歷 (U+6B74).

If a character variant is not included in Unicode 7.0, type the standard character instead. In addition, mark it by <g></g>.

Example:

A standard character 鼯 (U+5602) replacing an obscure non-unicode character “  ” in the original text.
<g>鼯</g>

If there is an unknown character in the text, i.e. a character variant which is readable but where you cannot identify the standard character, add it to the numbered list of unknown characters. For each character that is unknown use a question mark enclosed by <g></g>.

Example:
Three unknown or illegible characters
<g>???</g>

Special Symbols

Name	Symbol	Unicode	Substitute	Unicode
LARGE CIRCLE	◯	U+25EF		
HORIZONTAL ELLIPSIS	…	U+2026		
BOX DRAW. DOUBLE HORIZONTAL	═	U+2550		
BOX DRAW. HEAVY HORIZONTAL	━	U+2501		
BOX DRAW. HEAVY VERTICAL	┃	U+2503		
IDEOGRAPHIC SPACE		U+3000		
LESS-THAN	<	U+003C	〈	U+3008
GREATER-THAN	>	U+003E	〉	U+3009

The special characters “<” and “>” are reserved for the markup. Should these appear inside the text, please substitute them with the characters from this list.

Type Styles

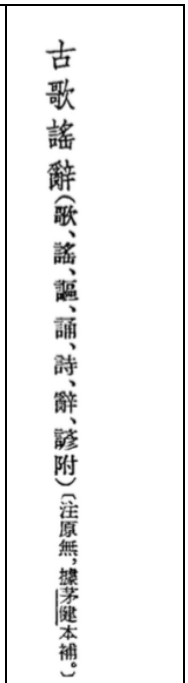
Small Characters

Strings of small characters, usually used for commentaries, are marked by <hi> </hi>.

Example:

Heading with two nested uses of small script.

```
<head>古歌謠辭
   <hi>（歌、謠、謳、誦、詩、辭、諺附）
      <hi>〔注原無，據<line>茅建</line>本補。〕
</hi>
   </hi>
</head><lb/>
```



Underlinings

A single line next to characters is marked by <line> </line>. A dotted line next to characters is marked by <del> </del>. A wavy line next to characters is marked by <mod> </mod>.
Please note: In old texts, the lines are to the right of characters. In modern texts,
the lines may also be to the left of characters. The position to the left or right is not encoded.
As with small characters above, a list of multiple

Example:
Different underlines in text.

[...] 後<line>太原</line><line>郭茂倩</line>輯<mod>樂府</mod>白卷, [...]	後太原郭茂倩輯樂府百卷,	[...] 此無他，~~無所以然之理以實乎其中故也~~。大凡一題到 [...]	此無他，無所以然之理以貫乎其中故也。大凡一題到

Please Note: In the example above there are two single lines and hence two <line></line> elements. Dots on the other hand, only require one pair of <del></del> elements for the whole string of characters.

List of All Tags

section	tag	name
Page Breaks	<pb/>	page break
Running Head	<fw type="header"></fw> <fw type="footer"></fw>	frame work
Text Blocks		division
Headings	<head></head>	heading
Paragraphs		paragraph
Line Breaks	<lb/>	line break
Front and Back matter	<text></text> <group></group> <front></front> <body></body> <back></back>	text group of texts front matter body back matter
Lists	<list> <item></item> </list>	lists with items
Tables	<table row=”1” col=”1”> <row> <cell></cell> </row> </table>	table with 1 row, 1 column, and no head.
Character variants and unknown characters	<g></g>	character variants, or unknown characters
Type Styles	<line></line> <mod></mod> <del></del>	single line wavy line dotted line
Small Characters	<hi></hi>	small type / commentary